Pre-miRNA Folding Through Context-Free Grammar Parsing and the Identification of miRNA Using a Feedforward Neural Network

Viktor Prypoten, Sean Gribben, Xavier Pellow, Reena Zelenkova, David Helmerson, Joshua Nibbs, Boris Deletic, Nathan Di Pierro, Andrew Harrison, Linda McIver, Sonika Tyagi (Monash University Faculty of Bioinformatics & John Monash Science School)

INTRODUCTION

There is currently no efficient way to accurately identify miRNA or predict the structure of miRNA in an ab-initio manner. The purpose of this work is to provide a framework which allows for efficient identification of mature miRNA and folding of pre-miRNA using a feedforward neural network (FFNN) and probabilistic context-free grammar (PCFG) parsing, respectively. The FFNN interprets and provides a prediction of the likelihood, expressed by a probability, of the input being miRNA.

BACKGROUND INFORMATION

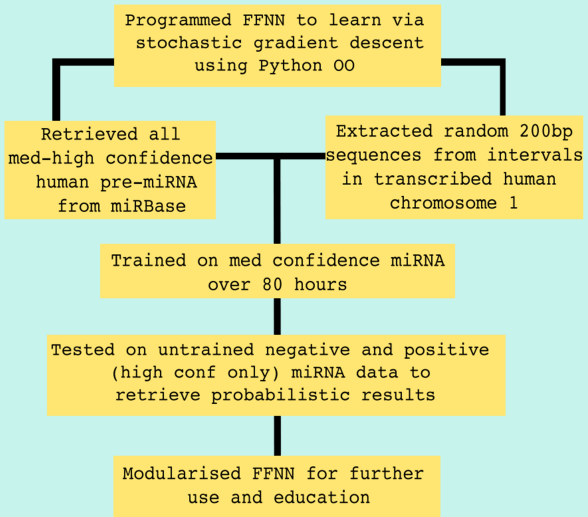
miRNA (micro RNA) serves as a mechanism employed by the cell to silence and/or slow down the expression of protein, via degradation of the pre-mRNA and mRNA through complementary binding and attachment to nucleases. The folding properties of pre-miRNA largely determine the activity of the enzymes involved in these steps, hence the miRNA as a therapeutic tool is deemed powerful due to its ability to simultaneously act on multiple genes due to their non-specific binding patterns.

Figure 1. Example of folded miRNA

FEED FORWARD NEURAL NETWORK (FFNN)

The intent of the FFNN was to interpret an input strand of miRNA and output an integer between 0 and 1 indicating the probability of its authenticity as miRNA. This drastically increases the efficiency of identification and inflates the source of information concerning miRNA by removing the need for ab initio evaluation. This could possibly lead to advancements in miRNA therapeutics and further clinical applications.

METHOD for FFNN:



Dimensions of the Network:

200 input neurons, 2 deep layers each of 16 neurons and 1 output neuron.

The FFNN was trained over 80 hours and achieved the following accuracies after training:

Sequence Pass Rate Against Neg/Pos miRNA Data	
Negative Data (random sequences from chromosome 1)	75%
Positive Data (high confidence miRNA sequences)	74%

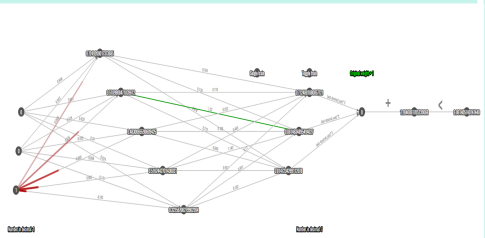
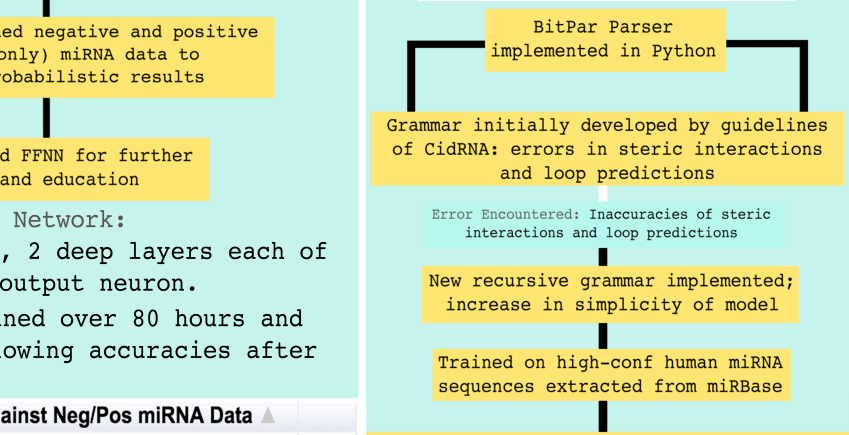


Figure 2. Demonstrative diagram of FFNN learning through stochastic gradient descent

PROBABILISTIC CONEXT-FREE GRAMMAR

PCFG (Probabilistic Context-Free Grammar) is a construct that describes the possibilities of arrangements (be it of words or miRNA), and the statistical weighting behind each permutation to determine the most likely arrangement. These arrangements are given as trees, which then can be converted to diagram structures.

METHOD for PCFG:



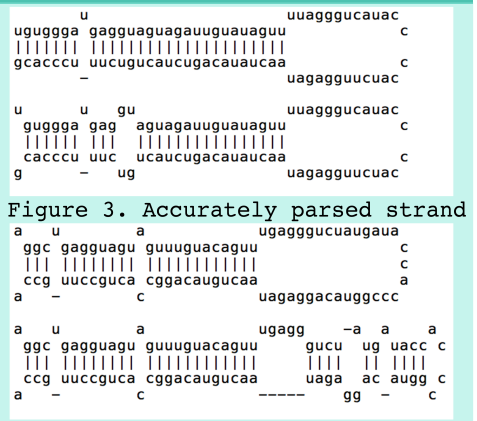


Figure 3. Accurately parsed strand

Figure 4. Comparison of strand indicating common parsing errors

CONCLUSION

The results indicate patterns in miRNA folding and positioning, hence two future pathways could reasonably be suggested. One would be to improve accuracy - through free energy calculations and addition of weighting parameters in the PCFG and restructuring and fine tuning of the FFNN the current accuracy could be improved. Alternatively, the programs could be linked together to directly funnel predicted miRNA strands to then be accordingly folded, and as a result be able to process and predict large chunks of currently sequenced genomes.

MONASH University